DOCUMENT RESUME

ED 206 723 TM 810 633

AUTHOR Eignor, Daniel R.; Hambleton, Ronald K.
TITLE Effects of Test Length and Advancement Score on Several Criterion-Referenced Test Reliability and Validity Indices. Laboratory of Psychometric and Evaluation Research Report No. 86.
INSTITUTION Massachusetts Univ., Amherst. School of Education.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
PUB DATE Jul 79
NOTE 37p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, 1979).

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Comparative Analysis; Computer Assisted Testing; Criterion Referenced Tests; *Cutting Scores; *Mastery Tests; Mathematical Models; Simulation; Test Construction; Test Format; *Test Reliability; *Test Validity
IDENTIFIERS *Binomial Error Model; *Test Length

ABSTRACT

The purpose of the investigation was to obtain some relationships among (1) test lengths, (2) shape of domain-score distributions, (3) advancement scores, and (4) several criterion-referenced test score reliability and validity indices. The study was conducted using computer simulation methods. The values of variables under study were set to be typical of those often used or obtained in practice. The reliability and validity indices (decision consistency, kappa, decision accuracy, predictive validity, and efficiency) are among the most useful indices for criterion-referenced test developers and evaluators. Practical guidelines are offered for using results obtained from the investigation to determine suitable lengths of criterion-referenced tests in specific assessment situations. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED206723

7/11/79

Effects of Test Length and Advancement Score on Several Criterion-Referenced Test Reliability and Validity Indices[1,2,3]

Daniel R. Eignor
Educational Testing Service

and

Ronald K. Hambleton
University of Massachusetts, Amherst

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

A primary concern of individuals using test scores is that the scores be both reliable and valid. While the best approach for assessing test score reliability and validity will depend on the particular situation, it is well-known that there is a relationship between the length of a test, the advancement score, and the reliability and validity of the test scores. Test scores with better psychometric properties (i.e., more reliable and valid) are obtained from longer tests.

For norm-referenced tests, the relationship of test length to reliability can be expressed by the Spearman-Brown formula. Also, formulas exist that relate norm-referenced test length to test score validity. However, because these formulas are based upon a correlational approach to reliability and validity, they are not very useful with criterion-referenced tests when the intent of the criterion-referenced test is to produce scores for making mastery/non-mastery

TM 810 633

[1]The project reported herein was performed persuant to a grant from the National Institute of Education, Department of Health, Education, and Welfare. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education and no official endorsement by the National Institute of Education should be inferred.

[2]Laboratory of Psychometric and Evaluative Research Report No. 86. Amherst, MA: School of Education, University of Massachusetts, 1979.

[3]Paper presented at the annual meeting of NCME, San Francisco, 1979.

decisions (Hambleton, Swaminathan, Algina, & Coulson, 1978). What is often of interest to users of criterion-referenced tests is information concerning the consistency of mastery/non-mastery decisions for some group of examinees across a retest administration or across a parallel-form administration. Also, there is usually considerable interest in the extent of agreement between mastery/non-mastery decisions based on a criterion-referenced test and the "true" mastery states of a group of examinees (sometimes called "decision accuracy"). (The "true" mastery state of an examinee is the one he/she should be assigned to, based on the amount of knowledge or skill he/she possesses relative to the objective or competency under investigation.) These two situations described above correspond to one paradigm for viewing the psychometric concepts of criterion-referenced test score reliability and validity, respectively (Millman, 1974; Popham & Husek, 1969).

Hambleton et al. (1978) distinguished between two uses of criterion-referenced test scores, domain score estimation and allocation of examinees to mastery states. For the first use, the test length relationship to reliability can be derived, and may be summarized by the well-known item sampling model (Lord and Novick, 1968). It is for the other major use of criterion-referenced test scores, mastery state determination, that necessary technical developments are in short supply. Little research has been done that directly explores the relationships of test length and advancement scores to criterion-referenced test score reliability and validity when the scores are used for assigning examinees to mastery states.

What research has been done has focused either (1) on procedures for determining reliability of examinee assignments to mastery states (Hambleton & Novick, 1973; Swaminathan, Hambleton, & Algina, 1974; Huynh, 1976; Subkoviak, 1976, 1978a, 1978b; Marshall & Haertel, 1976; Algina & Noe, 1978) or (2) on procedures for the determination of test length that minimizes misclassification errors (Millman, 1973; Novick & Lewis, 1974; Fhaner, 1974; Wilcox, 1976, 1977). The research reported in this paper is directed toward linking together these two areas of research and providing useful results for test practitioners to enable them to determine test lengths to fit the situations in which their tests will be used.

Specifically, the purpose of the study was two-fold:

1. To report the relationships between test lengths and several reliability and validity indices for a fixed cut-off score (80%) in five domain score distributions.

2. To report the relationships between advancement scores and several reliability and validity indices for several test lengths in five domain score distributions.

The study was carried out using computer simulation methods. The one major advantage of this approach is that it is possible "to know" examinee domain scores and their "true" mastery states. Such information permits one to compare examinee estimated domain scores and assigned mastery states, based on test results, with domain scores and true mastery states. A summary of such comparisons addresses the validity of the particular set of test scores under investigation.

Research Design

Terminology

Test length refers to the number of test items that are used to measure examinee performance on a particular objective. A domain score for an examinee (denoted, π_i) is the proportion of items in the domain of items measuring an objective that the individual can answer correctly. A cut-off score (denoted, π_o) on the domain score scale [0, 1] is used to separate examinees into two true mastery states.

Since all items in the domain of items defined by an objective cannot usually be administered to examinees for the purpose of assessing their domain scores or assigning them to mastery states, a sample of test items is chosen. Estimated domain score (denoted, $\hat{\pi}_i$) is the proportion of items that an examinee answers correctly of the items measuring an objective included in a test. An advancement score (denoted, $\hat{\pi}_o$) is the proportion of items measuring an objective on a test deemed necessary for an individual to answer correctly to be classified as a master.

In using an examinee's test score to determine his/her true mastery status, two types of classification errors can result. A false-positive error occurs when an examinee is estimated to be a master when his/her true status is non-master; a false-negative error occurs when an examinee is estimated to be a non-master when his/her true status is master.

Variables Under Study

(a) Test Model

Both the binomial and compound binomial models were used to simulate examinee item response data. While criterion-referenced test

data have often been assumed to fit the binomial model, Lord (1965) and Wilcox (1976, 1977) have suggested that the compound binomial model may be more appropriate. In the binomial model it is assumed that the probability of a correct item response for an examinee is the same across all items on a test. Or, to say it in another way, the assumption is made that all items are equally difficult (for that examinee). In the compound binomial model, it is assumed that the probability of a correct item response for an examinee varies from one item to the next in a test. The latter assumption is considerably more plausible, but investigations that have utilized both models (for instance, Subkoviak, 1976) have demonstrated different, but not very much different, results from the use of the two models.

(b) Prior Distributions

For the binomial model, either a user-supplied or a beta prior distribution of domain scores was specified and 200 examinee domain scores were generated. Two different methods were used to generate examinee domain scores under the assumption of the binomial test model. In one method, a test developer specifies his/her beliefs about the percentage of examinees located in different intervals (ten were used in the study, .00 to .10, .11 to .20, and so on) on the domain score scale [0,1]. Once the required number of examinees is specified and using a random number generator, a distribution of examinee domain scores can be produced which closely approximates the specified distribution. This method is especially easy to use. A second method involves the specification of parameters of a beta distribution representing a test developer's prior beliefs about the

distribution of domain scores (Novick & Jackson, 1974). It is then possible to obtain a random sample of examinee domain scores from the specified distribution.

Domain scores for use with the compound binomial model were generated from a normal distribution (mean = 0, standard deviation = 1) and then rescaled (by a linear transformation) to the interval [0,1]. This step and others done with the compound binomial model were carried out with the aid of computer program DATGEN (Hambleton & Rovinelli, 1973). The program has been used by many researchers to generate logistic test model data.

Additional details on the five domain score distributions used in the study are reported in Table 1 and Figure 1.

(c) Advancement Scores

In addressing the first purpose of the study, advancement scores were always set exactly equal to the chosen cut-off score of .80. This was possible because of the test lengths under consideration.

In the second part of the study, for two test lengths, advancement scores were varied with the same test data sets to determine the influence of advancement score placement on indices of test score reliability and validity.

(d) Test Lengths

Test lengths of 5, 10, 15, 20, and 40 were considered in this particular study. Many other test lengths (and advancement scores) were considered by Eignor (1979) in a similar study tc this one.

Table 1

Description of the Five Domain Score Distributions

Distribution	Test Model	Skewness	Domain Score Distribution Description
1	Binomial	Moderate Negative	(a) Mode is slightly below the cut-off score (.80). (b) Range of scores is [.11, 1.00]. (c) About 50% are on the interval [.60, .80] and 80% on the interval [.50 to .90].
2	Binomial	High Negative	(a) Leptokurtic distribution with the mode above the cut-off score (.80). (b) Range of scores is [.60, 1.00] with about 80% of the scores on the interval [.80, 1.00].
3	Binomial	Very High Negative	(a) Mode is above the cut-off score. (b) 50% on the interval [.00, .79] and 50% on the interval [.80, 1.00]. (c) Substantial variation of scores.
4	Compound Binomial	Moderate Negative	(a) Mode is close to the cut-off score. (b) Wide range of domain scores [.00, 1.00]. (c) 50% on the interval [.00, .79] and 50% on the interval [.80, 1.00]. (d) Flatter distribution than either (1) or (2).
5	Compound Binomial	None	(a) An almost rectangular distribution on the interval [.20, .90] for domain scores with fewer of them below .20 and above .90.

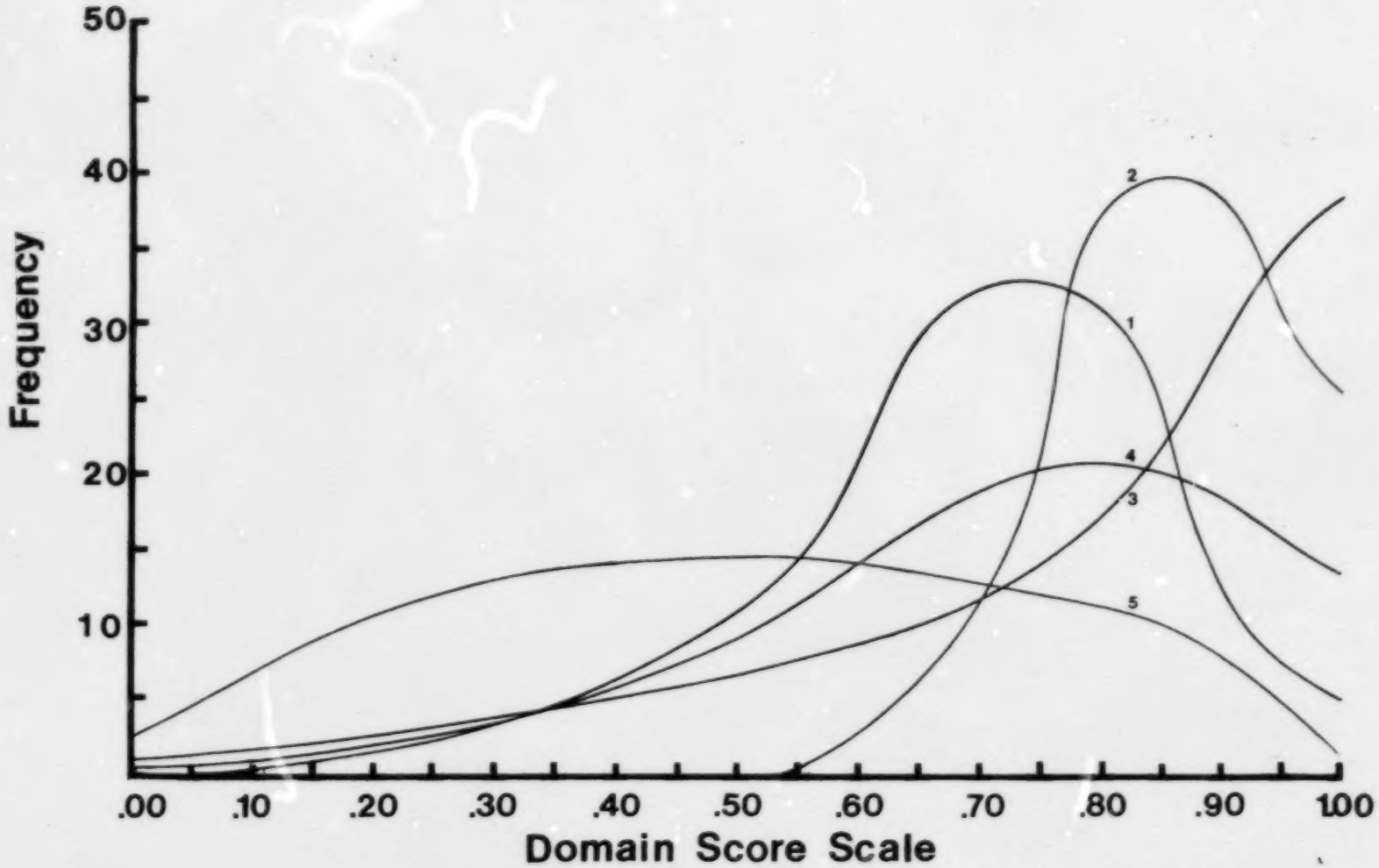

Figure 1. Graphical representation of five domain score distributions.

Reliability and Validity Indices

A number of relevant indices of test score reliability and validity were used in the study. The two diagrams below will facilitate a discussion of them.

Diagram One

Test Results-Occasion Two

Test Results-Occasion One		NM	M
	NM	p_{00}	p_{01}
	M	p_{10}	p_{11}

Diagram Two

Criterion Measure

Test Results		NM	M
	NM	p_{00}	p_{01}
	M	p_{10}	p_{11}

(M = Mastery status; NM = Non-Mastery status)

The contingency table in diagram one shows the proportion of examinees falling in the four possible combinations of mastery state assignments based on parallel-form (or test-retest) administrations of a set of test items measuring an objective included in a criterion-referenced test. A criterion measure is substituted for a parallel-form of the criterion-referenced test in diagram two.

Two reliability indices are derivable from data reported in Diagram One:

1. Decision Consistency

$$DC = \sum_{k=0}^{1} p_{kk}$$ (Hambleton & Novick, 1973)

2. Kappa

$$\kappa = \frac{DC-CA}{1-CA}$$ (Swaminathan, Hambleton, & Algina, 1974)

where CA (chance agreement) = $\sum_{k=0}^{1} p_{k\cdot}\, p_{\cdot k}$

and $p_{0\cdot}$, $p_{1\cdot}$, and $p_{\cdot 0}$, $p_{\cdot 1}$ are the respective marginal proportions for the first and second test administrations.

There are three derivable validity indices from Diagram Two:

3. Decision Accuracy

$$DA = \sum_{k=0}^{1} p_{kk}$$ (Hambleton and Novick, 1973)

4. Predictive Validity (the Pearson correlation between decisions based on the criterion-referenced test and the criterion measure) (Berk, 1976)

5. Efficiency

$$E = \frac{\sum_{i=1}^{N} (\pi_i-\pi_o)\ \text{Sign}\ (\hat{\pi}_i-\hat{\pi}_o)}{\sum_{i=1}^{N} |\pi_i-\pi_o|}$$ (Livingston, 1978)

where π_o is a cut-off score defined on the domain score scale, $\hat{\pi}_o$ is an advancement score, π_i is the domain score for examinee i and $\hat{\pi}_i$ is the estimated domain score for examinee i.

All of the statistics are well-known and commonly used in criterion-referenced testing practice except for the last one (and this is at least partially due to its newness). Essentially, efficiency is a measure of how accurately a criterion-referenced test and associated

advancement score result in the assignment of examinees to mastery states that are in agreement with decisions based on a criterion measure. Also, the loss in efficiency due to misclassifying examinees (false-positive, and false-negative errors) is linearly related to the difference between an examinee's level of performance on the criterion measure and the criterion measure cut-off score. Livingston's efficiency statistic does not address directly the validity of mastery classifications. The index was included in the study because it provided an alternate but potentially useful framework for viewing criterion-referenced test score validity.

Data Generation

The process of generating examinee item scores and test scores and summary statistics on 200 examinees for various sets of testing conditions was completed as follows:

1. One of the domain score distributions from Table 1 with accompanying test model (binomial or compound binomial) was selected.

2. Examinee domain scores were generated and examinees with domain scores equal to or above .80 were assigned to a mastery state on the criterion measure. All other examinees were assigned to a non-mastery state.

3. For the particular test length under consideration, examinee domain score estimates were generated. For the binomial test model, this was done by setting the probability of a correct response for each item equal to the examinee's domain score. By generating random numbers uniformly distributed on the interval [0, 1], it was possible to simulate the examinee's test item performance. This step was repeated to produce a second set of item scores for each examinee. The two sets of examinee item scores obtained on each examinee served as a basis for assessing test score reliability for the group of examinees under investigation. The initial set of item scores for each examinee was used in the validity portion of the study.

   For the compound binomial model, "item characteristic curves" were generated (see an example in Figure 2). From Figure 2 it is clearly seen that the probability of correct answers varies not only from one examinee to another but also for the same examinee from one item to another. Once probabilities for answering items for a given examinee were obtained, item scores via the use of a random number generator were obtained.

4. From the examinee item scores obtained in step 3, examinee test scores were obtained by summing the number of test items answered correctly.

5. Each examinee was assigned to a mastery state based on a comparison of his/her estimated domain score and the advancement score. Two assignments were made, one for each test administered.

6. The five summary statistics were calculated.

7. Steps 1 to 6 were repeated for each of five domain score distributions, and five test lengths (5, 10, 15, 20, and 40 test items). In addition, for two test lengths (5 and 10), the summary statistics were calculated for three advancement scores, one at 80%, and one below and the other one above 80%.

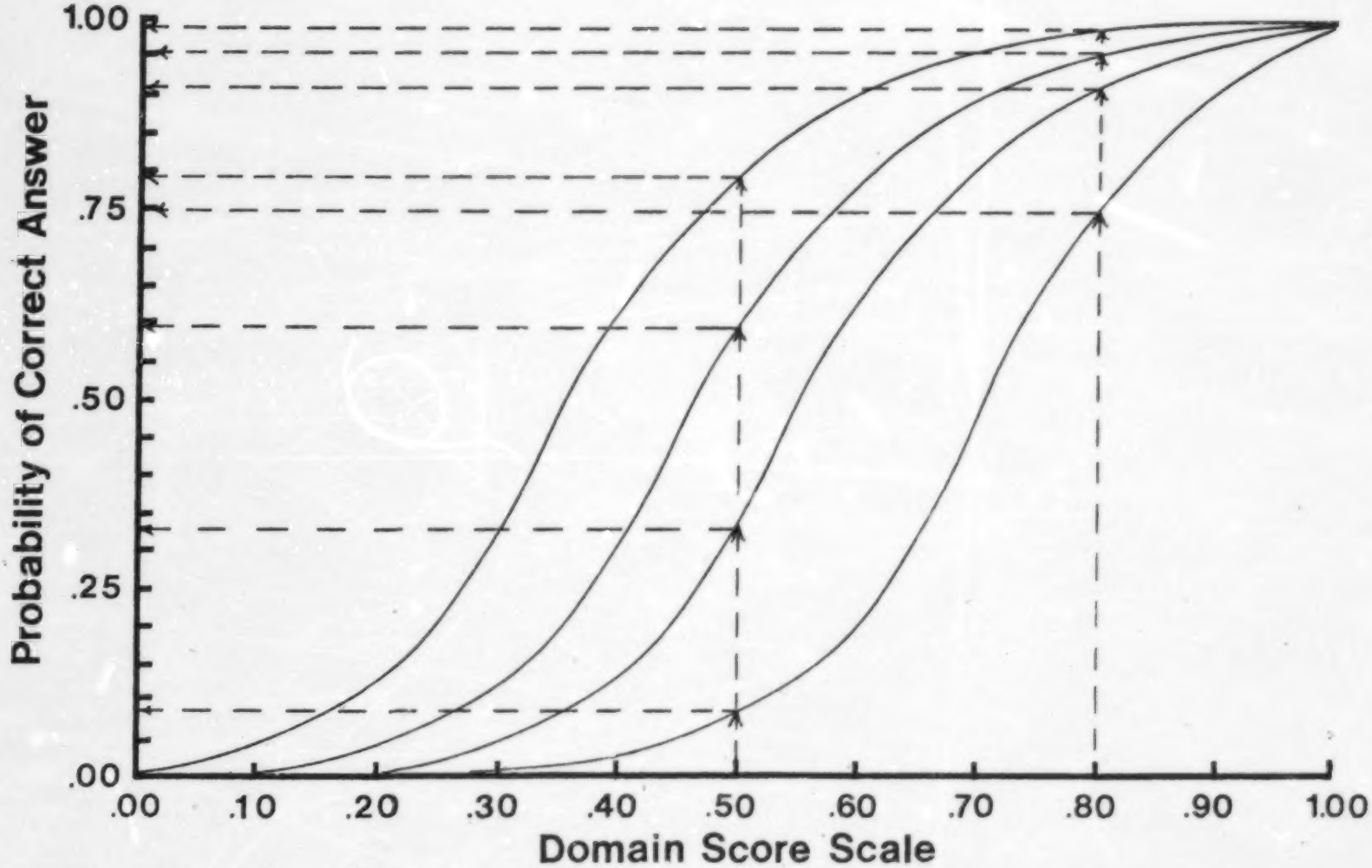

Figure 2. Item Characteristic Curves of Four Test Items and Probabilities of Correct Answers for Two Examinees

Results and Discussion

Effects of Test Length on Selected Test Score Reliability and Validity Indices

Figures 3 to 7 provide the relationships between test length and decision consistency, kappa, decision accuracy, predictive validity, and efficiency, respectively, for each of the five domain score distributions under consideration. In preparing the figures, statistical data were available for each of the domain score distributions at six test lengths: 0, 5, 10, 15, 20, and 40 items. Curves were drawn to be monotonically increasing, non-intersecting, and as close fitting to the data points as possible.

A number of observations and/or cautions concerning the use of Figures 3 to 7 are offered next:

1. Test score validity indices are lowest with homogeneous domain score distributions centered at or near the cut-off score. Domain score distribution one (and to a lesser extent) distribution two reflect this. The validity indices are highest when domain scores distributions are homogeneous and located far from a cut-off score. These findings have several implications:

   Shorter tests can be used when there is reason to believe that a group of examinees will do either very well or very poorly on a particular test. (Of course, if the prior belief about the distribution of domain scores is highly inaccurate, test score validity indices will be considerably lower than those predicted from the figures.)

2. Figures 3 to 7 apply to the case $\pi_0 = \hat{\pi}_0 = .80$. Such a situation is common in practice, but variations in cut-off scores and advancement scores from .80 will reduce the usefulness of the results reported in the figures.

3. Details for using the figures in test development work will be offered later in the paper. It suffices to say here that the more important figures are those connecting test length to the validity indices. After an initial determination of test length has been made through the consideration of a relevant index, Figures 3 and 4 can be used to predict test score reliability. If it is

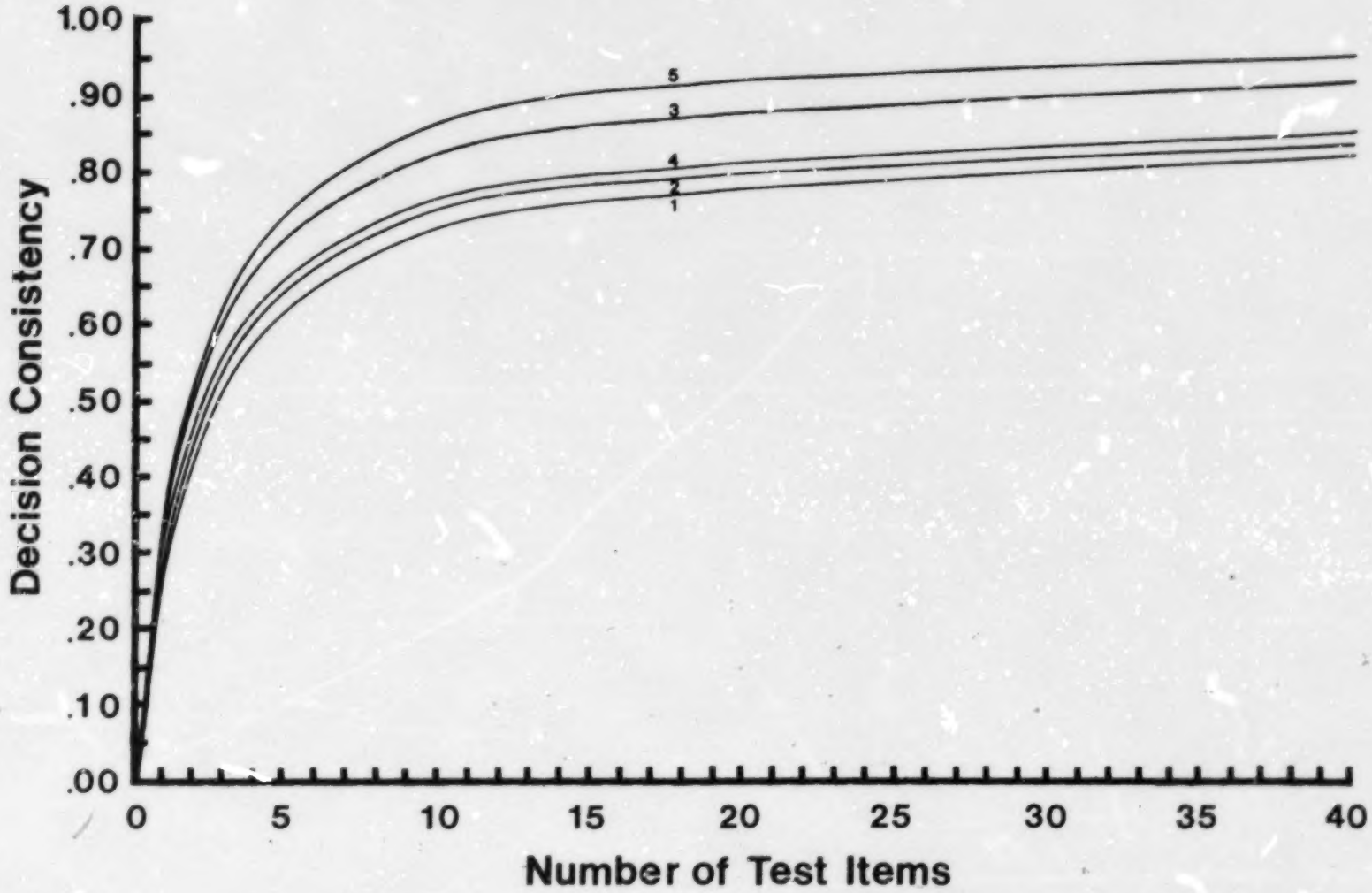

Figure 3. Relationship Between Decision Consistency and Test Length with Five Test Score Distributions

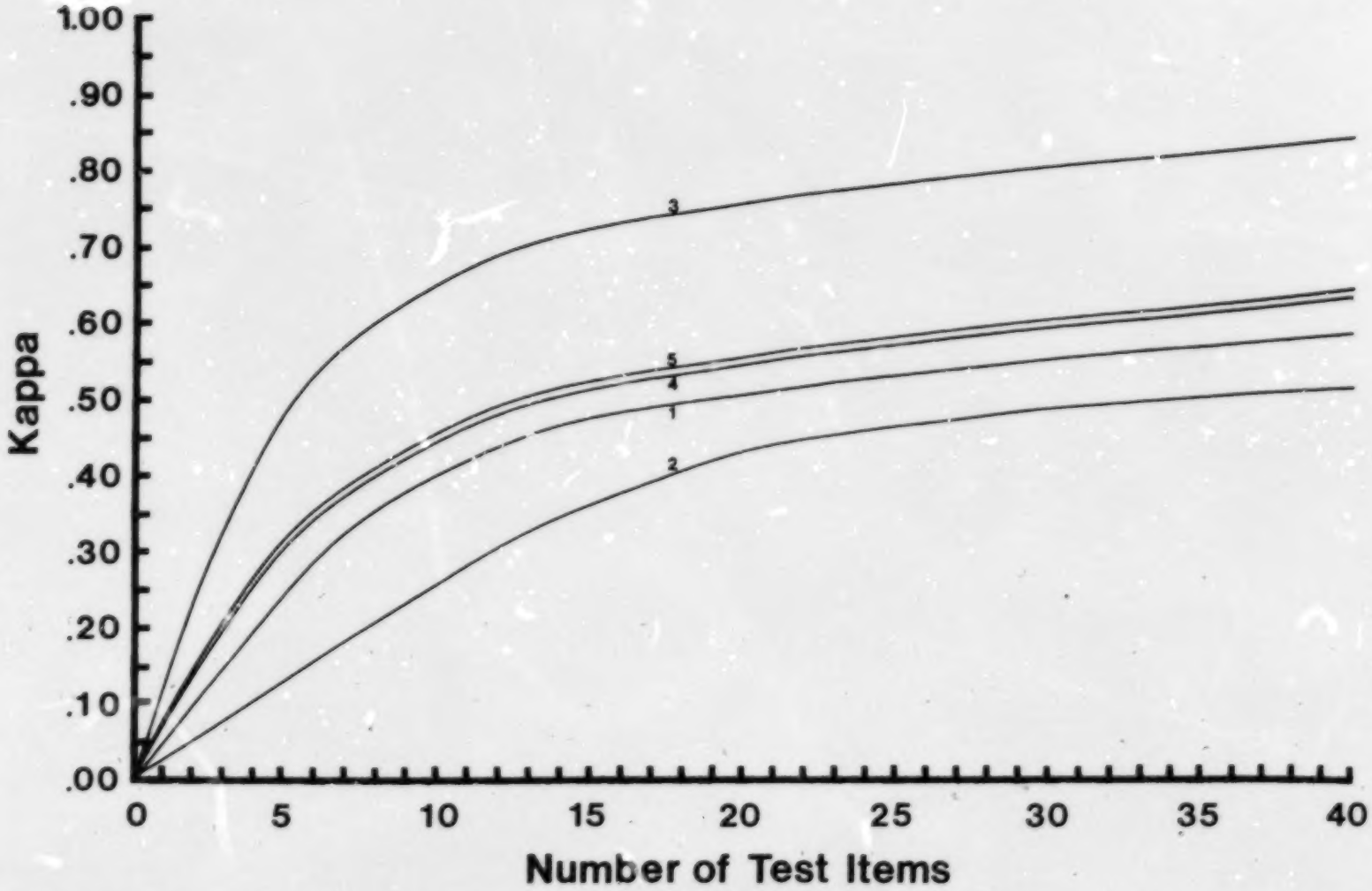

Figure 4. Relationship Between Kappa and Test Length with Five Test Score Distributions

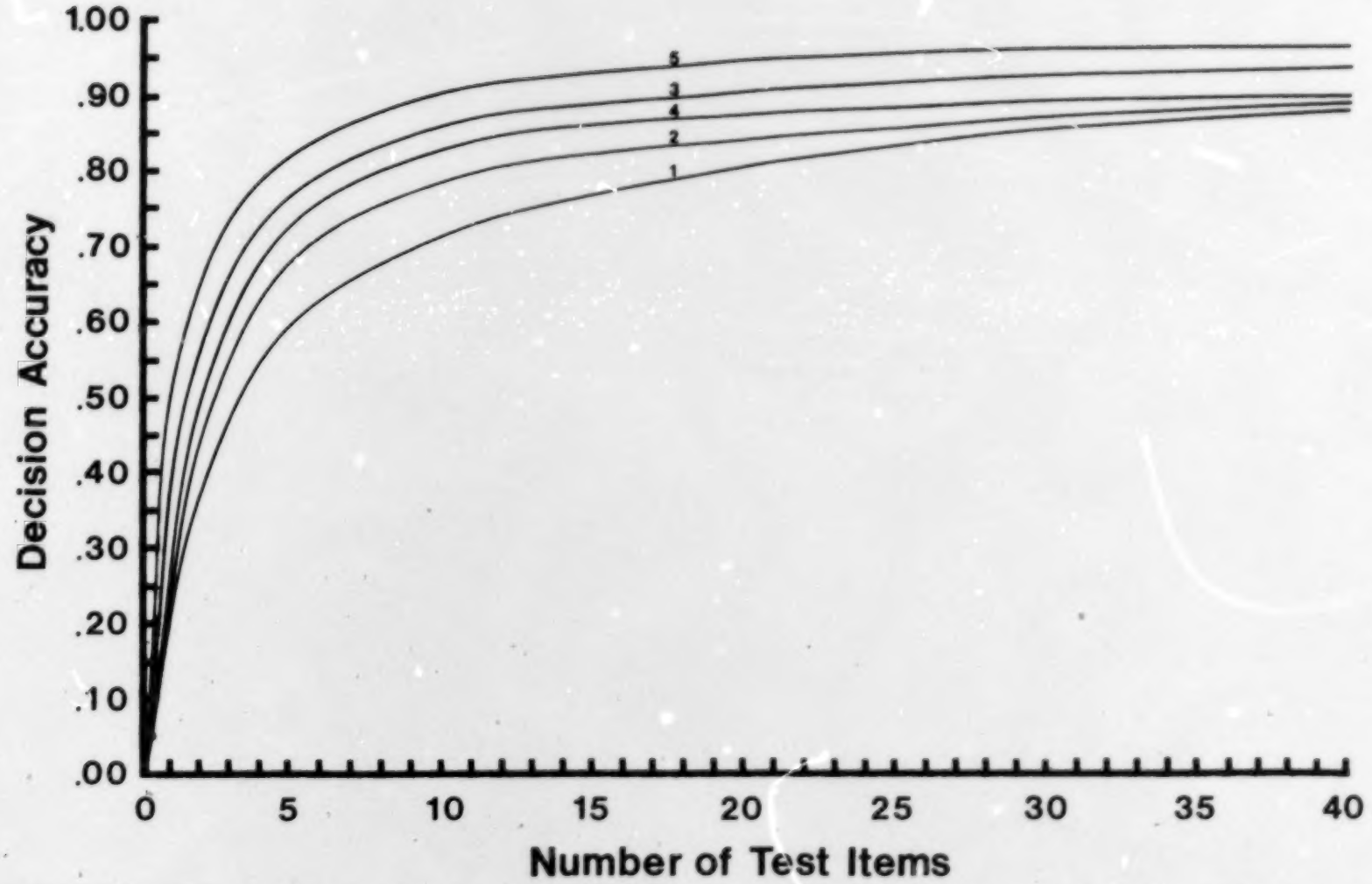

Figure 5. Relationship Between Decision Accuracy and Test Length with Five Test Score Distributions

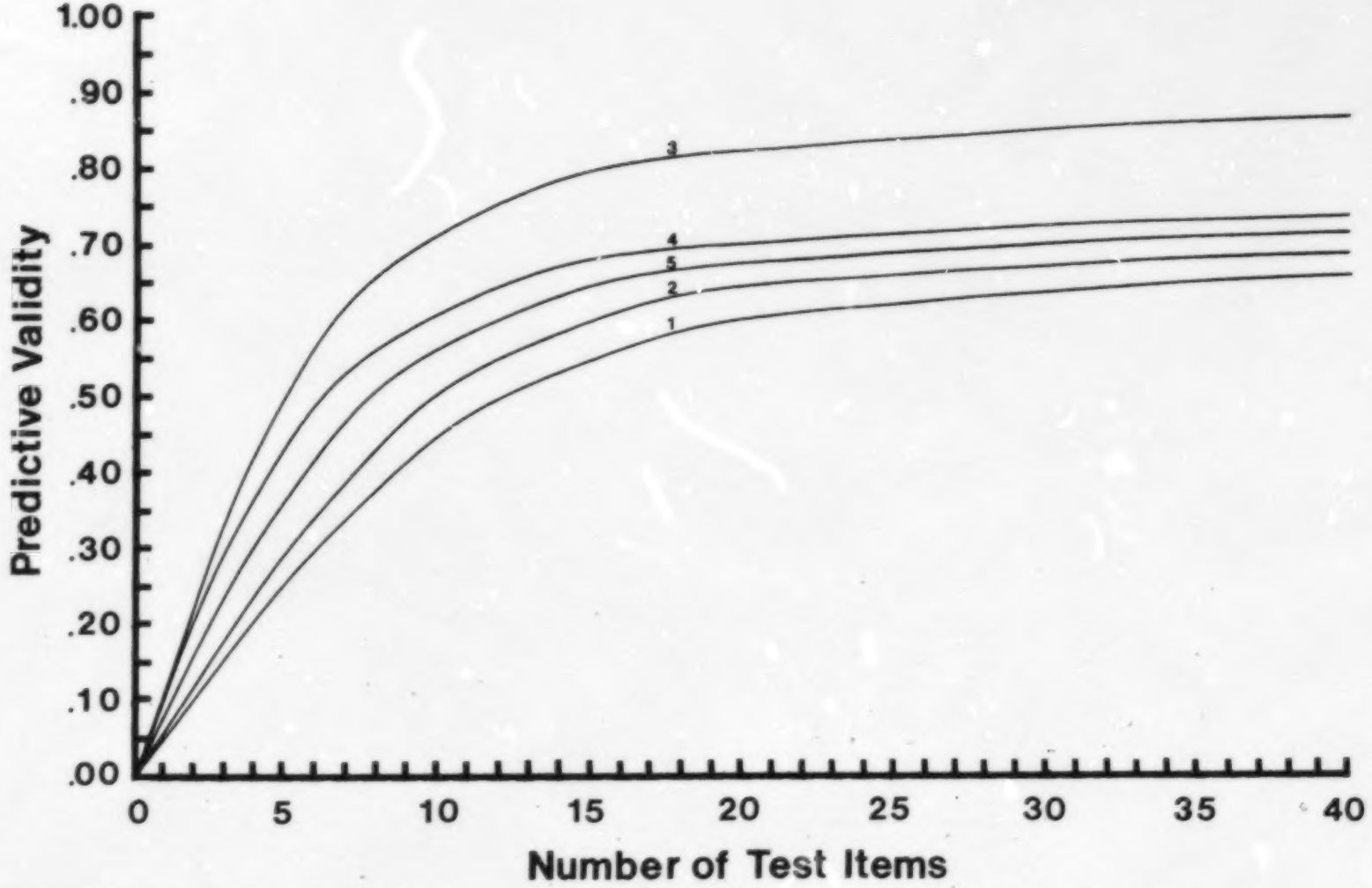

Figure 6. Relationship Between Predictive Validity and Test Length with Five Test Score Distributions

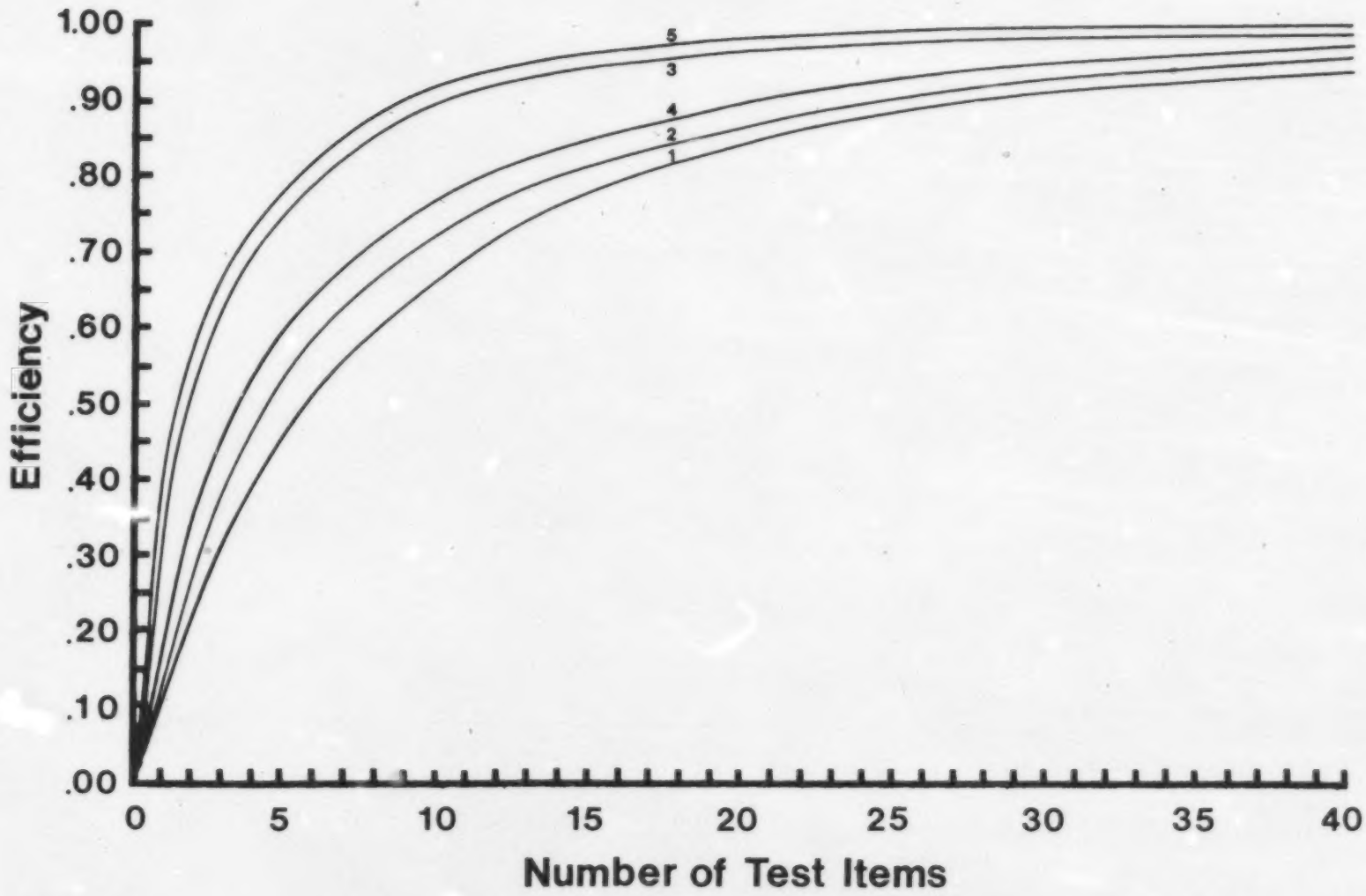

Figure 7. Relationship Between Efficiency and Test Length with Five Test Score Distributions

not high enough to meet some specified standard, the test plan must be revised to lengthen the required test.

Effects of Advancement Score on Test Score Reliability and Validity Indices

It is not always possible to set a cut-off score and an advancement score equal to the same value. Sometimes it is not even desirable to do so even when the opportunity is available. For example, if false-positive errors are considerably more serious than false-negative errors, a test user may choose to set a very high advancement score and thereby minimize the number of false-positive errors. Such an action, however, will influence test score reliability and validity indices.

Reliability and validity indices for two test lengths, three advancement scores, and five domain score distributions are reported in Table 2. A few comments are then offered to help in the interpretation of the results in the Table. Note, however, that because of sampling errors, not all of the results are consistent with the interpretations offered below.

Table 2

Effect of Advancement Score on Several Reliability and Validity Indices with Five Domain Score Distributions

Statistic	Test Length	Advancement Score	Domain Score Distribution 1	2	3	4	5
Decision Consistency	5	3	.72	.93	.84	.76	.71
	5	4	.64	.71	.76	.66	.71
	5	5	.74	.55	.76	.70	.87
	10	7	.73	.84	.80	.77	.73
	10	8	.74	.74	.81	.72	.86
	10	9	.77	.62	.84	.74	.89
Kappa	5	3	.22	.08	.58	.32	.41
	5	4	.28	.11	.49	.31	.35
	5	5	.24	.10	.49	.29	.34
	10	7	.47	.16	.51	.45	.30
	10	8	.46	.28	.62	.43	.47
	10	9	.33	.23	.67	.40	.40
Decision Accuracy	5	3	.43	.70	.72	.55	.62
	5	4	.60	.74	.82	.68	.76
	5	5	.83	.59	.82	.74	.88
	10	7	.56	.80	.75	.74	.78
	10	8	.69	.77	.87	.77	.90
	10	9	.83	.71	.89	.83	.95
Predictive Validity	5	3	.25	.09	.54	.36	.25
	5	4	.29	.32	.65	.40	.43
	5	5	.48	.22	.64	.48	.40
	10	7	.31	.38	.56	.54	.33
	10	8	.42	.51	.75	.55	.55
	10	9	.50	.42	.78	.58	.39
Efficiency	5	3	.02	.51	.55	.15	.51
	5	4	.45	.64	.78	.53	.75
	5	5	.82	.39	.83	.75	.93
	10	7	.43	.71	.81	.62	.79
	10	8	.66	.70	.89	.74	.93
	10	9	.83	.61	.93	.88	.97

Decision Consistency

As the advancement score is moved away from the center of a domain score distribution, decision consistency increases. This explains why for the 10-item test and distribution five, decision consistency is lowest (.73) at $\hat{\pi}_0 = .70$ and highest (.89) at $\hat{\pi}_0 = .90$. The mean of the distribution is in the region of .60. The reverse result is obtained with distribution two. The highest value (.84) is obtained at $\hat{\pi}_0 = .70$ and the lowest value (.62) is obtained at $\hat{\pi}_0 = .90$. The mean of distribution two is about .90. Since the mean of distribution four is close to .80, it is not surprising to observe the lowest value (.72) at $\hat{\pi}_0 = .80$ and higher values at $\hat{\pi}_0 = .70$ (.77) and at $\hat{\pi}_0 = .90$ (.74).

Kappa

While the results are not clear cut, it appears that the highest values of kappa are obtained when an advancement score is near the middle of a domain score distribution. Huynh (1976) noted a similar finding in some of his work.

Decision Accuracy

The value of decision accuracy is monotonically related to the distance between $\hat{\pi}_0$ and $\bar{\pi}$. The role that π_0 plays in the tabulated results is not readily apparent from the reported results. It is known that decision accuracy will be increased when π_0 and $\hat{\pi}_0$ are equal.

Predictive Validity

There do <u>not</u> appear to be any trends in the results.

Efficiency

The results here are identical to those reported for decision accuracy and the explanation is the same.

Using the Results to Determine Test Length

Many factors will have an influence on the test length which is finally selected:

1. The shape (essentially variability) of the domain score distribution (regardless of which statistic is chosen, it is clear from Figures 3 to 7 that the variability of the

domain score distribution has a considerable influence on the results). In general, higher indices are obtained with heterogeneous domain score distributions.

2. The placement of cut-off scores (in general, higher validity indices are obtained if π_0 and $\bar{\pi}$ are not too close).

3. The selection of advancement scores (has a complicated relationship to test length).

4. The desired level of one of the reliability and/or validity indices (the higher the desired value, the longer the required test must be).

Six steps are offered next for determining test length in particular testing situations:

1. Select a reliability or validity statistic of interest (this is usually "decision accuracy").

2. Set a cut-off score (if $\pi_0 = .80$, proceed through the remaining steps; if $\pi_0 \neq .80$, it will be necessary to generate additional results using the method described in the last section of this paper).

3. Set advancement scores corresponding to test lengths under consideration which are near .80 (if $\hat{\pi}_0 \doteq .80$ Figures 3 to 7 will provide usable results).

4. Specify a prior belief about the domain score distribution for the group of examinees who will be assessed. If conservative results are desired, it is best to work with homogeneous distributions centered around $\pi_0 = .80$.

5. Choose (a) or (b)

   (a) With the statistic identified in step 1, and a desired value of the statistic, use the proper figure and read off the corresponding test length from the curve corresponding to the domain score distribution selected in step 4.

   For example, suppose a test developer desired the decision accuracy statistic to equal .80 and the most likely domain score distribution is number 1. From Figure 5, the corresponding test length is 21 items.

   (b) With the reliability or validity statistic selected in step 1, and several test lengths of interest, find the corresponding values of the desired statistic for the

test lengths of interest. Select the test length which is acceptable for the testing situation.

6. Check "decision consistency" and/or "kappa" for the test length selected in step 5. (With the example in 5a above, the value is .75 for decision consistency.) If the value is too low for the intended purpose of the test, determine a value which is not, and read off the corresponding test length.

The values provided in the figures are only approximations. Still, they should be helpful to test developers who aspire to set their test lengths in a way which is not totally dependent on guess work.

Suggestions for Further Research and Development

Because of (1) the considerable importance of the topics under study in this paper, and (2) the paucity of practical research results, it is easy to suggest many directions for further work. For one, a computer program is needed into which a test developer can (a) provide a prior belief about the shape of a domain score distribution for some group of examinees to be tested, (b) select a test model (probably the binomial or the compound binomial), (c) select one or more reliability and validity indices of interest, and (d) select test lengths and advancement scores of interest. The output from the computer program would provide a basis for determining test length.

One of the spin-offs from this simulation study is the availability of a computer program that has some of the features mentioned above. It can be used by test practitioners to generate additional results to those reported in the paper. Practitioners must only (1) specify a prior belief about the distribution of domain scores, (2) suggest test lengths, cut-off scores, and advancement scores, and

(3) select either the binomial or compound binomial test model from which to simulate examinee item response data. Figures similar to those reported in this study can be quickly obtained. A write-up of our current computer program is in preparation.

A second area for additional work is in the area of "guidelines for interpreting reliability and validity indices." In the area of norm-referenced testing, even with a plethora of textbooks available and the training many people have had, there is still considerable confusion about the correct interpretations of reliability and validity indices. Because of the relative newness of the five statistics used in this study, it seems clear that if they are to have any value at all, increased effort must be given to training test developers in the use of these and other relevant statistics (see for example, Hambleton & Eignor, 1979).

Third, the validity of the relationships reported in Figures 3 to 7 among test length, cut-off scores, advancement scores, and domain score distributions, and five reliability and validity indices should be compared to existing results reported on real test data. In a limited way, some of the reliability results reported in this paper have been compared to results obtained with real data. The differences were small. Considerably more work of this type should be done. The reliability results would be particulalry easy to check. Only examinee item responses to large sets of test items keyed to objectives would be required. "Tests" of varying lengths can be drawn from an examinee-item pool of data keyed to a particular

objective, "parallel-forms" constructed, and various advancement scores considered. Via the method of purposive sampling of examinees, assuming the "pool" of examinees was heterogeneous and large enough, the influence of the shape of a domain score distribution on test score reliability can also be studied.

References

Algina, J., & Noe, M. J. A study of the accuracy of Subkoviak's single-administration estimate of the coefficient of agreement using two true-score estimates. Journal of Educational Measurement, 1978, 15, 101-110.

Berk, R. A. Determination of optimal cutting scores in criterion-referenced measurement. Journal of Experimental Education, 1976, 45, 4-9.

Eignor, D. R. Psychometric and methodological contributions to criterion-referenced testing technology. Unpublished doctoral dissertation, University of Massachusetts, 1979.

Fhaner, S. Item sampling and decision-making in achievement testing. British Journal of Mathematical and Statistical Psychology, 1974, 27, 172-175.

Hambleton, R. K., & Eignor, D. R. A practitioner's guide to criterion-referenced test development, validation, and test score usage. Laboratory of Psychometric and Evaluative Research Report No. 70. Amherst, MA: School of Education, University of Massachusetts, 1979. (2nd ed.)

Hambleton, R. K., & Novick, M. R. Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 1973, 10, 159-170.

Hambleton, R. K., & Rovinelli, R. A Fortran IV program for generating examinee response data from logistic test models. Behavioral Science, 1973, 17, 73-74.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, 1978, 48, 1-47.

Huynh, H. On the reliability of decisions in domain-referenced testing. Journal of Educational Measurement, 1976, 13, 253-264.

Livingston, S. A. Assessing the reliability of tests used to make pass/fail decisions. COPA Research Report. Princeton, NJ: Educational Testing Service, 1978.

Lord, F. M. A strong true-score theory, with applications. Psychometrika, 1965, 30, 239-270.

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading, MA: Addison-Wesley, 1968.

Marshall, J. L., & Haertel, E. H. The mean split-half coefficient of agreement: A single administration index of reliability for mastery tests. Unpublished manuscript, University of Wisconsin, 1976.

Millman, J. Passing scores and test lengths for domain-referenced measures. Review of Educational Research, 1973, 43, 205-216.

Millman, J. Criterion-referenced measurement. In W. J. Popham (Ed.), Evaluation in education: Current applications. Berkeley, CA: McCutchan Publishing Co., 1974.

Novick, M. R., & Jackson, P. H. Statistical methods for educational and psychological research. New York: McGraw-Hill, 1974.

Novick, M. R., & Lewis, C. Prescribing test length for criterion-referenced measurement. In C. W. Harris, M. C. Alkin, and W. J. Popham (Eds.), Problems in criterion-referenced measurement. CSE monograph series in evaluation, No. 3. Los Angeles: Center for the Study of Evaluation, Univesrity of California, 1974.

Popham, W. J., & Husek, T. R. Implications of criterion-referenced measurement. Journal of Educational Measurement, 1969, 6, 1-9.

Subkoviak, M. Estimating reliability from a single administration of a criterion-referenced test. Journal of Educational Measurement, 1976, 13, 265-275.

Subkoviak, M. J. Empirical investigation of procedures for estimating reliability for mastery tests. Journal of Educational Measurement, 1978, 15, 111-116. (a)

Subkoviak, M. J. The reliability of mastery classification decisions. Paper presented at the First Annual Johns Hopkins University National Symposium on Educational Research, Washington, October 27, 1978. (b)

Swaminathan, H., Hambleton, R. K., & Algina, J. Reliability of criterion-referenced tests: A decision-theoretic formulation. Journal of Educational Measurement, 1974, 11, 263-268.

Wilcox, R. A note on the length and passing score of a mastery test. Journal of Educational Statistics, 1976, 1, 359-364.

Wilcox, R. R. Estimating the likelihood of false-positive and false-negative decisions in mastery testing: An empirical Bayes approach. Journal of Educational Statistics, 1977, 2, 289-307.